Homework 8:
Corpora and Lexical Resources

Benjamin Roth, Marina Sedinkina
Symbolische Programmiersprache

Due: Thursday January 11, 2017, 16:00

In this exercise you will:

- implement a language guesser

Exercise 1: Language Guesser [6 points]

Implement a language guesser that determines the language it thinks the text is written in. The decision should base on the frequency of individual characters in each language. Take a look at the file `hw08_lang_guesser/model_lang.py`. In this exercise you will have to complete some methods to make it work.

This homework will be graded using unit tests by running: `python3 -m unittest -v hw08_lang_guesser/test_lang_guesser.py`

1. Complete the class method `build_language_models(self)`. This method should return a conditional frequency distribution where:
   - the languages are the conditions
   - the values are the lower case characters
2. Complete the class method `guess_language(self,language_model_cfd, text)`:
   - it should calculate the overall score of a given text based on the frequency of characters accessible by `language_model_cfd[language].freq(character)`.
   - it should return the most likely language for a given text according to the scores
3. The previous language guesser was based on the frequency of characters. English and German texts are difficult to distinguish with the given approach. In `model_bigram_lang.py` implement alternative class called `LangBigramModeler` which will have the same methods but based on the character bigrams. Discuss, which approach works better.

4. Once you have implemented all missing functionality, you can have a look at `guess_lang.py` to see how to use it in practice. Run the code with: `python3 -m hw08_lang_guesser.guess_lang`